IBM Applied Data Science Capstone Project

# Introduction

This report summarizes the findings in my IBM applied data science capstone project. The project’s intention was to answer the following question:

## "Which city would be best to open up a restaurant and what type of restaurant should one open?"

### Methodology

First stage of the project will focus on retrieving data from the data sources and clean the data so that they can be processed with Pandas data frame. Data was scrapped using BeautifulSoup4. Scrapped data usually contains unneeded characters and signs and should be cleaned.

Second stage of the project will focus on retrieving geolocation data from Foursquare and look for restaurants in the target city. The project will count the number of restaurants, broken down into categories. The project will determine a type of restaurant that is least present in the area.

Third stage of the project will seek for additional information that will reinforce the recommendation for a restaurant in the area from the above.

#### Data sources

* List of largest cities (<https://en.wikipedia.org/wiki/List_of_largest_cities)>,
* List of cities by GDP (<https://en.wikipedia.org/wiki/List_of_cities_by_GDP)>.

#### Target audience

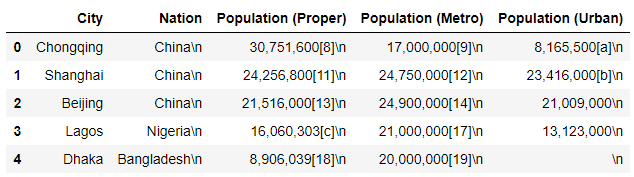
Someone looking to open up a restaurant globally.

#### Data sources

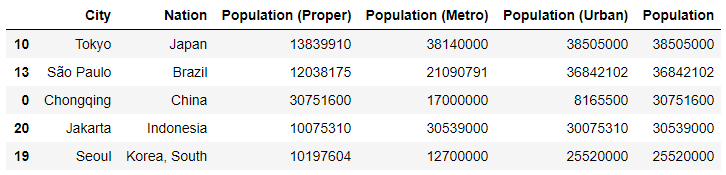
* Population, GDP of cities: Wikipedia
* Geolocation coordinate data: <https://simplemaps.com/data/world-cities>
* Place data: Foursquare

# data retrieval and cleaning

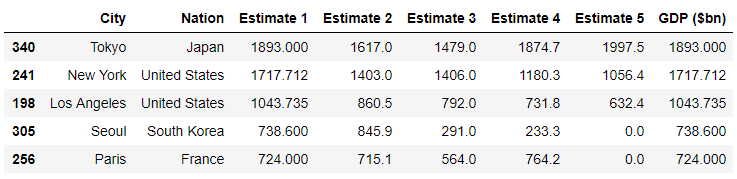
1. Data for population initially scrapped from Wikipedia had many unnecessary data which interfered with mathematical calculation.  
   The following table shows the uncleaned table

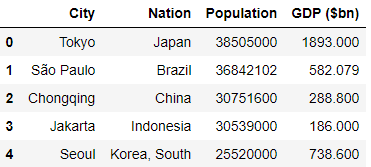


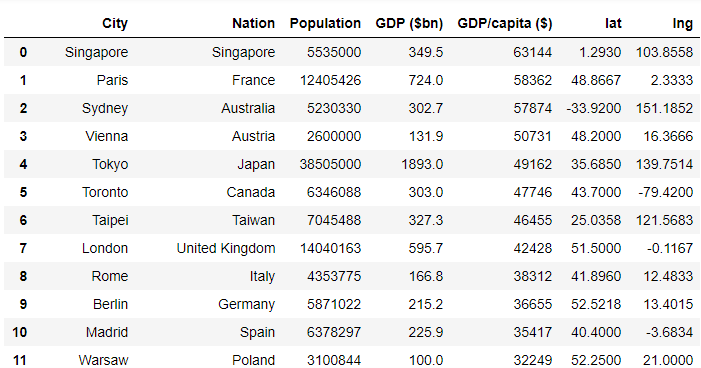
1. String modifications were applied to the data frame to remove contents in the parenthesis. Also the \n signs were also removed  
   After cleaning the data, notice that the columns that should have numbers, such as population data has been changed to type int.

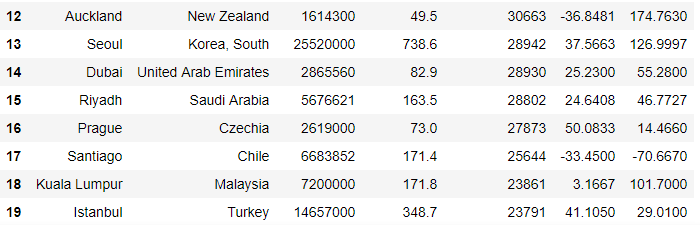


1. The same cleaning process was applied for the GDP data from Wikipedia.  
   The table below shows the dataframe.head() for the GDP data retrieved for the major cities.  
   Notice that there are multiple cities for a nation.



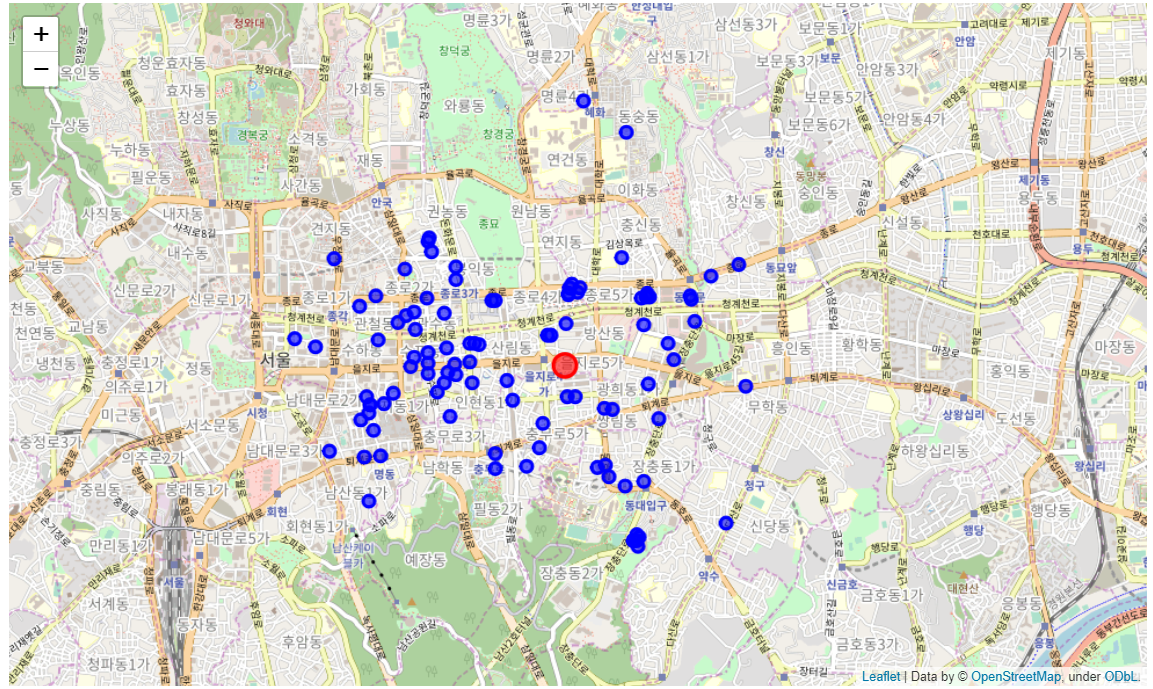
1. Before merging the data frame, city with the highest GDP was selected and the rest dropped from the data frame.  
   The below shows data frame after merging.  
     
   
2. In order to retrieve geospatial data, data was downloaded from <https://simplemaps.com/data/world-cities> due to problems with retrieving data from the Google platform. After merging data, columns lat and lng was added to the data frame for each of the major cities in the data frame.  
   The table below shows all the data retrieved (population, GDP, geospatial data) for the top 20 countries with the highest GDP/capita in US$.



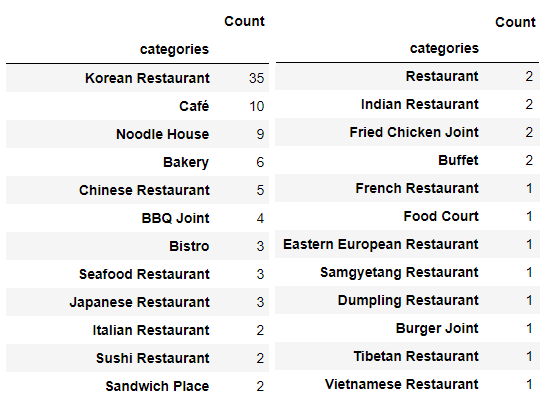


# Using Foursquare to retrieve venue information for a location

1. Seoul was selected arbitrary from the list of top 20 countries
2. Foursquare data retrieved for the Seoul coordinate of type “food”
3. The red circle depicts the coordinate passed onto Foursquare for retrieval. Due to restrictions, only around 50 venues are retrieved from each of Foursquare queries. The food venues retrieved from the query is marked as a blue circle in the map below.



1. The below shows the count of venues classified into type of restaurant.



# Initial result

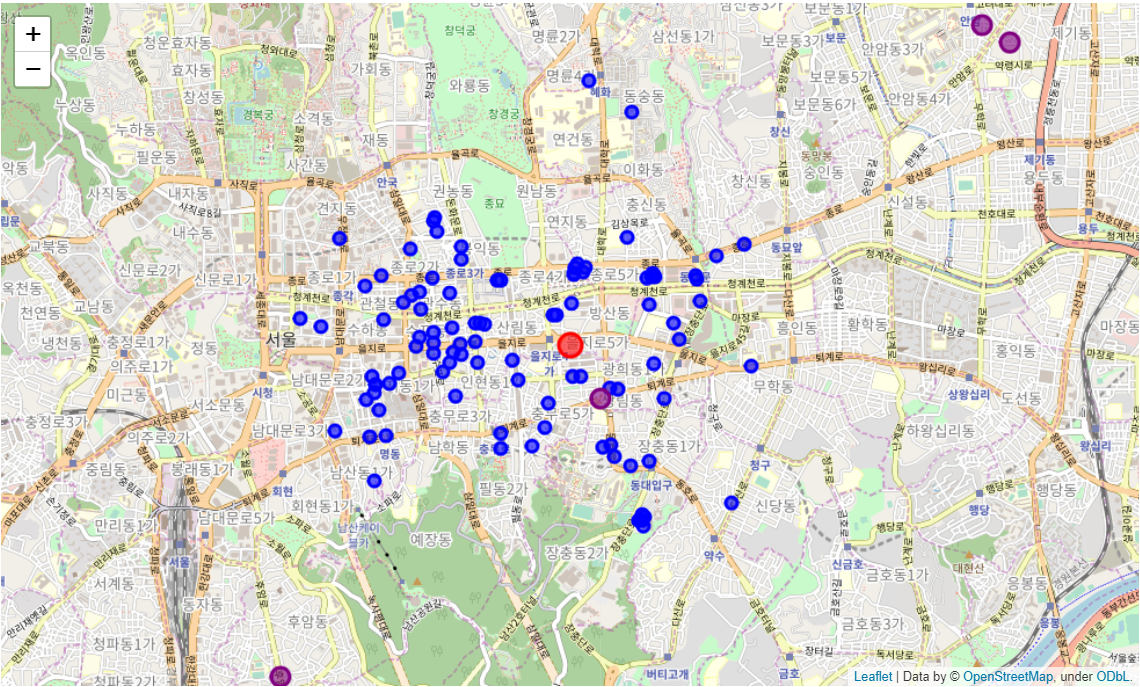
The initial thesis is that there are no restaurants related to Islam. The project makes the assumption that there should be demand for Halal foods if there are enough Islam population in the city. Research shows that there are around 100,000 Islams living within Seoul.

The idea is to see if there are Islam mosques in Seoul and locate them and see whether it would be feasible to have a Halal food restaurant within the target area.

# Result discussion

Query for Islam mosques from Foursquare shows that there are two Mosques within the immediate vicinity of the initial search area. For the larger area around Seoul, there are a total of five mosques that are within 30 minutes’ drive, which would be considered to be a feasible distance for customers.

The map below shows restaurants in blue circle and Islam mosques as purple dots. Red circle shows the central point for the search queries.



# conclusion

The project concludes that there is potential for prospective investors or entrepreneurs to open up a Halal food restaurant in Seoul in the norther district.